МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Физтех-школа физики и исследований им. Ландау

Математическая статистика конспект

Автор:
Яренков Александр Владимирович

Долгопрудный
20 мая 2024 г.

Definitions

Генеральная совокупность - то, что дробится на выборки (т.к. всех данных слишком много)

Выборка должна из себя представлять МОДЕЛЬ генеральной совокупности. Должна быть моделью. Тогда выборка называется РЕПРЕЗЕНТАТИВНОЙ.

Простая случайная выборка (simple random sample (SRS))

Стратифицированная выборка - разбиваем ген совокупность на РАЗЛИЧНЫЕ по своей природе страты (группы)

Групповая выборка - разбиваем ген совокупность на ПОХОЖИЕ по своей природе страты (группы)

Т.о. чем меньше выборка, тем больше отклонение среднего выборки от среднего генеральной совокупности

SE - Standart Error - стандартная ошибка генеральной совокупности

ESE - Estimate Standart Error - стандартная ошибка выборки делённая на $\sqrt{n}$, где n - количество наблюдений в выборке

Выборочное среднее - среднее по выборке

Распределение выборочных средних - распределение, показывающее какие значения принимает среднее значение выборки из генеральной совокупности при многократном случайном выборе разных выборок. Согласно ЦПТ при количестве выборок стремящемся к бесконечности, мы получим нормальное распределение со средним значением генеральной совокупности и дисперсией в $\sqrt{n}$ раз меньше дисперсии генеральной совокупности

Гистограмма работает с численными данными, а столбчатая диаграмма с категориальными

Число x является a-квантилем набора данных <=> ($a \cdot 100\%$ данных <= x) И ($100\% - a \cdot 100\%$ данных >= x) Т.о. a - квантиль, a * 100% - перцентиль

Ковариация - мера совместной изменчивости двух величин

Коэффициент корреляции Пирсона - мера ЛИНЕЙНОЙ зависимости между двумя величинами. Поэтому в случае нелинейных зависимостей его применять не стоит. Также его не стоит применять при наличии выбросов, т.к. "под капотом"он считается как мат ожидание, а значит чувствителен к выбросам. Этот коэффициент помогает узнать связанность величин, но не помогает узнать что является следствием другого. Возможно, вообще связанность данных двух факторов связана с наличием некоего третьего фактора, влияющего на исходные два. То есть он лишь указывает на наличие ЛИНЕЙНОЙ зависимости, но не утверждает что она обязательно есть.

Бинаризация - преобразование числовой переменной в категориальную методом разделения на интервалы

Нулевая гипотеза (H_0) - гипотеза об отсутствии различий/изменений

Альтернативная гипотеза (H_1) - гипотеза о наличии различий/изменений

Ошибка первого рода - отклонение верной нулевой гипотезы. Вероятность совершить эту ошибку $-\alpha$. Или p-value - максимально допустимая вероятность совершить ошибку первого рода

Ошибка второго рода - принятие неверной нулевой гипотезы. Вероятность совершить эту ошибку $-\beta$. Мощность статистического теста равна $1 - \beta$.

Tests

Z-test

УСЛОВИЯ ПРИМЕНИМОСТИ: нормальное распределение случайной величины

Величина $z = \frac{x-Ex}{\sigma}$, где x - среднее выборки, Ex - среднее генеральной совокупности, σ - стандартная ошибка среднего, называется z-статистика. Статистика здесь в смысле некоторого числа, получаемого по данной формуле. Расчёт z-статистики и определение по ней возможность отклонить нулевую гипотезу и есть Z-тест. Например, если $z = 3$, то это означает что среднее выборки находится на расстоянии 3σ от среднего генеральной совокупности. При значении $\alpha = 0.05$ это означает, что у

нас достаточно оснований отклонить нулевую гипотезу, так как это значение α это всё равно что 1.96σ

УСЛОВИЯ ПРИМЕНИМОСТИ: знание дисперсии генеральной совокупности

T-test

0.0.1 Одновыборочный

УСЛОВИЯ ПРИМЕНИМОСТИ: нормальное распределение выборочных средних

$H_0 :< x >= \mu$ Рассчитываем значение $t = \frac{<x>-\mu}{ESE}$ Имеет n-1 степеней свободы (число независимых случайных величин) При бесконечном n является стандартным нормальным распределением. При n > 30 очень близко к нормальному. Пик у него ниже, а хвосты, соответственно выше. ПОЗВОЛЯЕТ УЗНАТЬ P-VALUE БЕЗ ЗНАНИЯ ДИСПЕРСИИ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

0.0.2 двухвыборочный

УСЛОВИЯ ПРИМЕНИМОСТИ: независимость средних по выборкам И нормальные распределения выборочных средних

$H_0 :< x_1 >= \mu_1 < x_2 >= \mu_2$ или $< x_1 > - < x_2 >= \mu_1 - \mu_2$. Рассчитываем значение $t = \frac{(<x_1>-<x_2>)-(\mu_1-\mu_2)}{ESE}$

Если распределение не является нормальным, то можно подробить генеральную совокупность на выборки случайным образом и тогда в этих выборках, возможно, будет приближённо наблюдаться нормальное распределение.

Двухпропорционный Z-test

$H_0 : p_1 = p_2$, где p_1 и p_2 - пропорции/доли

УСЛОВИЯ ПРИМЕНИМОСТИ: ...

Рассчитываем значение $z = \frac{(p_1-p_2)}{\sqrt{p(1-p)(1/n_1+1/n_2)}}$, где n_1 - количество данных в выборке 1, n_2 - количество данных в выборке 2, а $p = \frac{p_1 n_1+p_2 n_2}{n_1+n_2}$

U-test Манна-Уитни

УСЛОВИЯ ПРИМЕНИМОСТИ: Независимость выборок. В каждой из выборок должно быть не менее 3 значений признака. Либо в одной выборке 2 значения, но во второй тогда не менее 5.

Для использования u-test'а нужно составить единый ранжированный по возрастанию ряд из двух выборок ($iin 1, 2$). Если есть одинаковые числа, то в качестве ранга берётся среднее арифметическое рангов одинаковых чисел.

n_i - количество наблюдений в выборке i

R_i - сумма рангов в выборке i $U = \min n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$

Далее по таблице для избранного уровня статистической значимости определить критическое значение критерия для данных выборок 1 и 2. Если наше значение U меньше, чем критическое, то есть статистически значимая разница, иначе - нет.